Apprentissage d'un module stochastique de compréhension de la parole

Hélène Maynard, Fabrice Lefèvre

Groupe Traitement du Langage Parlé, LIMSI-CNRS, FRANCE
{Helene.Maynard,Fabrice.Lefevre}@limsi.fr

ABSTRACT

The need for human expertise in the development of a speech understanding system can be greatly reduced by the use of stochastic techniques. However corpus-based techniques require the annotation of large amounts of training data. Manual semantic annotation of such corpora is tedious, expensive, and subject to inconsistencies. In order to decrease the development cost, this work investigates the performance of the understanding module with two parameters: the influence of the training corpus size and the use of automatically annotated data.

1. INTRODUCTION

Le rôle du module de compréhension dans un système de dialogue oral est d'extraire le sens littéral d'une requête de l'utilisateur. L'approche stochastique a permis de réduire fortement le recours à l'expertise humaine lors du développement du module de compréhension [4, 6, 5]. Cette approche, fondée sur un apprentissage automatique, nécessite néanmoins l'annotation manuelle d'une quantité importante de données d'entraînement. L'annotation sémantique des corpus est un travail fastidieux et coûteux, qui peut être sujet à erreurs. La diminution du coût lié à l'annotation des corpus constitue donc un enjeu important.

Nous présentons tout d'abord le développement du modèle stochastique de compréhension sur des données collectées avec le système ARISE du LIMSI. Les performances obtenues par le modèle sont comparées à celles obtenues par la grammaire de cas actuellement utilisée dans les systèmes de dialogue du LIMSI. Puis deux expériences sont menées avec pour objectif de réduire le coût de développement du modèle : (1) l'évaluation de l'impact de la quantité de données manuellement annotées sur les performances du modèle et (2) la mise en oeuvre d'une procédure automatique pour l'annotation des données en terme de marqueurs de concepts. Le coût de l'annotation manuelle pourra alors être diminué d'une part en sélectionnant une quantité de données adaptée et d'autre part en limitant l'annotation aux concepts de base.

2. LA TÂCHE ARISE

Le système de dialogue ARISE du LIMSI [2] est destiné à la réservation de billets de trains par téléphone. Il permet aux utilisateurs d'obtenir des renseignements sur des trajets (horaires, types des trains,...) mais aussi sur des informations telles que les tarifs, les possibilités de réduction et les prestations (bar, couchettes,...).

2.1. Représentation sémantique

Une représentation sémantique spécifique à la tâche ARISE a été utilisée. La faisabilité des processus d'annotation et d'évaluation dépend grandement du choix de cette représentation. La représentation concept/valeur (CVR) choisie repose sur un *dictionnaire de concepts* contenant une liste de 64 concepts représentant la tâche, avec pour chaque concept la liste des valeurs qui peuvent lui être associées. Un exemple de représentation CVR d'un énoncé est donné dans la dernière ligne de la figure 1. Les valeurs sont des nombres, des noms propres ou des formes normalisées d'expressions qui sont synonymes pour la tâche. Par exemple pour le concept `plage`, les expressions *dans la matinée*, *le matin* ou *avant midi* sont normalisées en une forme unique `matin`. Une information de modalité (affirmative ou négative) est attribuée à chaque concept. Un énoncé est donc représenté par une liste de triplets [mode, concept, valeur normalisée] (ex : [`+/ville: Roissy`]). L'ordre des triplets de la représentation respecte l'ordre d'apparition des concepts dans la phrase.

2.2. Compréhension par grammaire de cas

Dans le système ARISE actuel, le processus de compréhension est fondé sur une analyse par grammaire de cas. A partir de la suite de mots fournie par le module de reconnaissance de la parole, un prétraitement sémantico-lexical est effectué. Il permet d'obtenir une forme normalisée des énoncés dépendante de la tâche. L'analyse par grammaire de cas proprement dite est ensuite réalisée sur les valeurs normalisées pour obtenir la représentation sémantique de l'énoncé. Un ensemble de mots-clés permet de sélectionner la structure de cas appropriée. Il est complémenté par un ensemble de marqueurs de cas permettant d'exprimer des contraintes syntactiques. Dans la phrase *de Paris à Marseille*, la préposition *de* indique que *Paris* est une ville de départ et la préposition *à* indique que *Marseille* est une ville d'arrivée. Le jeu de marqueurs comprend des pré- et des post-marqueurs, des marqueurs adjacents ou éloignés et des anti-marqueurs. L'analyse par la grammaire de cas est donc effectuée en sélectionnant par mots-clés la structure de cas appropriée et en instanciant les valeurs des cas à l'aide des marqueurs.

Les deux traitements de l'analyse utilisent un ensemble de règles de réécritures et une grammaire décrites manuellement. Leur mise au point et leur maintenance requièrent une expertise humaine importante.

Requête	*dans la matinée*	*et*	*c'est*	*pas*	*Croisic*	*c'est*	*Roissy*
Reconnaissance	*dans la matinée*	*et*		*pas*	*Croisic*		*Roissy*
Décodage conceptuel	`+/plage-dep`	`+/null`		`-/m:mode`	`-/ville`		`+/ville`
Normalisation	`matin`				`Croisic`		`Roissy`

Représentation sémantique (CVR)	`+/plage-dep` `matin`
	`-/ville` `Croisic`
	`+/ville` `Roissy`						

Figure 1: *Exemple de décodage sémantique d'un énoncé.*

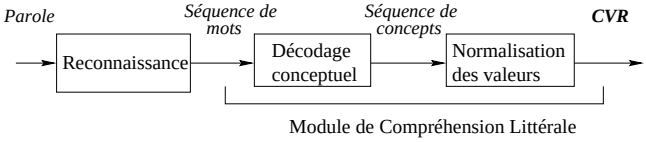

Figure 2: *Schéma de la compréhension stochastique.*

3. LA COMPRÉHENSION STOCHASTIQUE

La compréhension stochastique est fondée sur un décodage conceptuel. Son principe est de déterminer la séquence d'unités sémantiques (*concepts*) qui va représenter au mieux le sens d'un énoncé en faisant l'hypothèse qu'il existe une correspondance séquentielle entre les séquences de mots et les séquences de concepts [4].

Soit $W = w_1 w_2 \dots w_L$ la séquence de mots composant la phrase, le décodage conceptuel consiste à trouver la séquence de concepts $\hat{C}$ qui maximise la probabilité *a posteriori* :

$$\hat{C} = \arg\max_C \Pr(C|W) \quad (1)$$

A l'aide de la formule de Bayes, l'équation (1) peut être réécrite :

$$\hat{C} = \arg\max_C \Pr(W|C)\Pr(C) \quad (2)$$

Le terme $\Pr(W|C)$ est estimé par des probabilités n-grammes de mots conditionnées au concept du mot i :

$$\Pr(W|C) \simeq \prod_{i=1}^{N} \Pr(w_i|w_{i-1}, \dots, w_{i-n}, c_i)$$

et le terme $\Pr(C)$ est estimé en terme de probabilités m-grammes de concepts :

$$\Pr(C) \simeq \prod_{i=1}^{N} \Pr(c_i|c_{i-1}, \dots, c_{i-m})$$

Dans les expériences décrites ci-dessous, nous avons utilisé des bigrammes de concepts (*i.e.* $m = 1$) et des probabilités conditionnelles des mots sur les concepts (*i.e.* $n = 0$).

Ces probabilités sont apprises sur un corpus annoté manuellement en concepts et en marqueurs de concepts. Les marqueurs, comparables aux marqueurs de cas, permettent de lever certaines ambiguïtés sur les concepts de l'énoncé : `-/m:mode` associé au mot *pas* dans la figure 1 permet, par exemple, de déterminer la modalité de [`-/ville: Croisic`]. Par ailleurs, le concept `null` est affecté aux mots n'apportant aucune information sémantique dans l'énoncé (*et* dans la figure 1). En tenant compte de la modalité, 170 concepts et marqueurs de concepts sont pris en compte par le module de compréhension stochastique.

Table 1: *Description des corpus : nombre de requêtes, de mots et de concepts présents dans les ensembles d'apprentissage, de développement, et de test. Les taux d'erreurs de reconnaissance sont donnés pour les ensembles de développement et de test.*

	Appr.	Dev.	Test
Nb requêtes	14582	400	496
Nb mots	72380	2261	2880
Nb concepts	44812	708	923
Taux err. mots	-	13,4%	14,3%

Le schéma du processus de compréhension complet est donné dans la figure 2 et un exemple avec les résultats de chaque étape du processus est donné dans la figure 1. Le module de reconnaissance transforme le signal acoustique en la suite de mots la plus probable (figure 1, 2ème ligne). Aucune transformation *a priori* des mots n'est effectuée, excepté pour les mots uniquement associés au concept `null` dans le corpus d'apprentissage. Ces mots n'ayant pas de sens pour la tâche, tels que *euh*, *ah*, ou encore *je*, sont retirés des énoncés avant le processus de décodage. Le décodage de Viterbi fournit la séquence de concepts alignée sur la suite de mots (figure 1, 3ème ligne). Après avoir retiré tous les triplets associés au concept `null` et aux marqueurs de concept, une étape de normalisation transforme les valeurs associées aux concepts en valeurs prévues par le dictionnaire de concepts (figure 1, 4ème ligne). Dans l'exemple, la suite de mots *dans la matinée* associée au concept `plage-dep` est transformée en la valeur normalisée `matin`. On obtient ainsi la liste des triplets formant la représentation sémantique CVR de l'énoncé traité.

4. DESCRIPTION DES CORPUS

L'ensemble d'apprentissage utilisé pour nos expériences contient 14 582 énoncés. Ces requêtes sont issues du corpus ARISE du LIMSI comprenant plus de 10k dialogues d'utilisateurs interagissant avec le système. Ce corpus a été annoté manuellement en terme de concepts [3]. Le nombre moyen de mots par phrases est de 5. Le nombre total de concepts de l'ensemble d'apprentissage est de 44 812, avec une moyenne de 3 concepts de base par phrase.

Un ensemble de développement de 400 phrases a été utilisé pour valider la procédure d'évaluation. L'évaluation est effectuée sur un ensemble de test de 496 requêtes sélectionnées aléatoirement dans la partie non utilisée du corpus ARISE. Un processus itératif a

été mis en œuvre pour établir les représentations CVR de référence. Afin de ne pas favoriser une approche en particulier, les approches par règles et stochastique ont été appliquées aux transcriptions manuelles des énoncés. La représentation CVR proposée par l'approche stochastique a été ensuite corrigée manuellement. Puis les CVR résultants ont été utilisés comme référence pour évaluer le résultat de l'approche par règles et ont conduit le cas échéant à de nouvelles corrections manuelles. Les caractéristiques des ensembles d'apprentissage, de développement et de test sont résumées dans la table 1.

Le système de reconnaissance de parole a un lexique de reconnaissance d'environ 4000 mots incluant plus de 3000 noms de gares. Le taux d'erreurs sur les mots est de 14,3% pour le corpus de test (13,4% pour le corpus de développement). L'ensemble des énoncés du corpus a été transcrit manuellement.

5. EXPÉRIENCES

Trois séries d'expériences sont rapportées. La première constitue une validation de l'approche stochastique par une comparaison de ses performances avec celles de l'approche par règles. Nous évaluons ensuite l'impact de la quantité de données utilisées lors de l'apprentissage du modèle sur ses performances. Enfin, une procédure permettant de limiter l'annotation manuelle de l'ensemble d'apprentissage aux seuls concepts de base de la tâche est proposée et évaluée. Le taux d'erreurs de compréhension est mesuré en terme de suppressions, insertions et substitutions sur les triplets [mode, concept, valeur] de la CVR.

5.1. Comparaison avec la grammaire de cas

Afin de tester l'efficacité de l'approche stochastique, ses performances ont été comparées avec celles de l'approche par règles [1]. Les résultats sont donnés dans la table 2. Lorsqu'elle est appliquée sur des transcriptions manuelles d'énoncés connus, l'approche par règles est très performante : sur les transcriptions manuelles de l'ensemble de développement, elle obtient un taux d'erreurs de compréhension de 2,1% contre 7,8% pour l'approche stochastique. La différence de performance entre les deux approches est significativement réduite dans le cas de transcriptions automatiques. La perte de performance due aux erreurs de reconnaissance se traduit par une multiplication par 6 du taux d'erreurs de compréhension pour l'approche par règles, contre 2 pour l'approche stochastique. Les performances de l'approche stochastique sont relativement stables entre les corpus de développement et de test, contrairement à celles de la grammaire de cas. Finalement, on observe que les deux approches obtiennent des performances comparables lorqu'elles sont confrontées à des énoncés inconnus : le taux d'erreurs de compréhension est d'environ 9% sur les transcriptions manuelles et de 19% sur les transcriptions automatiques.

5.2. Influence de la taille du corpus d'apprentissage

L'annotation sémantique du corpus d'apprentissage constitue un travail fastidieux et coûteux. Afin de réduire le coût de la compréhension stochastique, l'impact de la quantité de données d'apprentissage sur les performances du modèle stochastique a été étudié.

Table 2: *Taux d'erreurs de compréhension pour les approches par grammaire de cas et stochastique sur les ensembles de développement et de test. Transcriptions exactes (*Manuelle*) et issues du module de reconnaissance (*Auto*).*

	Dev.		Test	
Approche	*Manuelle*	*Auto.*	*Manuelle*	*Auto.*
Grammaire de cas	2,1	13,2	9,2	19,8
Stochastique	7,8	16,6	9,4	19,1

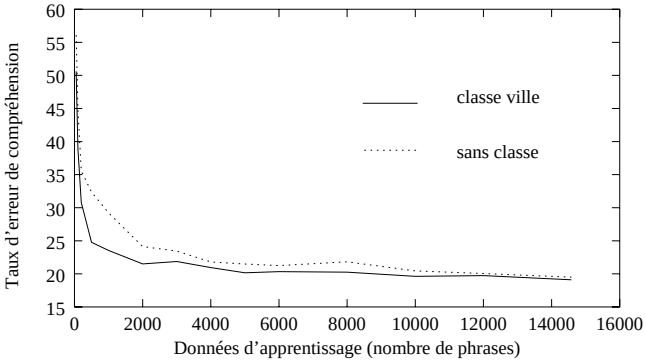

Figure 3: *Taux d'erreur de compréhension en fonction du nombre d'énoncés utilisés pour l'apprentissage avec (trait plein) et sans (trait pointillé) classe lexicale pour représenter les villes.*

La variation du taux d'erreurs de compréhension en fonction du nombre d'énoncés utilisés pour l'apprentissage est donnée dans la figure 3. Ces résultats ont été obtenus avec les transcriptions automatiques des énoncés de test et dans deux conditions d'apprentissage : avec (trait plein) et sans (trait pointillé) classe lexicale pour représenter les villes. On observe sur la figure 3 que le taux d'erreurs de compréhension décroît très rapidement et se stabilise ensuite. Ce phénomène est amplifié par l'utilisation d'une classe lexicale. Dans ce cas, le taux d'erreurs se stabilise à partir de 2 000 requêtes. Sans la classe lexicale, le double de requêtes est nécessaire pour atteindre un niveau de performance équivalent. Au-delà de 11 000 requêtes, l'influence de la classe devient négligeable .

Du point de vue du développement, on peut déduire de ces observations que l'annotation du corpus d'apprentissage peut être limitée à quelques milliers d'énoncés sans perte importante de performance : le taux d'erreurs est de 21,5% avec 2000 requêtes, contre 19,1% avec l'ensemble d'apprentissage complet. Ce résultat pose néanmoins le problème de la complexité du modèle utilisé dès lors qu'une quantité importante de données n'est manifestement pas prise en compte par le modèle. Il conviendra d'étudier plus précisement si cet effet est imputable à une redondance des données ou aux limites du modèle.

5.3. Utilisation de marqueurs de concepts automatiques

Dans les expériences décrites précédemment, l'ensemble des marqueurs de concepts a été établi *a priori* et se fonde sur une expertise humaine. L'annotation suppose alors une segmentation complète des phrases en concepts de base et en marqueurs de concepts. Nous proposons une procédure de sélection automatique des marqueurs qui permet de limiter l'annotation manuelle aux seuls concepts de base de la tâche.

	$p=s=0$	$p=1, s=0$	$p=2, s=0$	$p=2, s=1$
je souhaiterais	-	-	-	-
aller	-	-	**+/ville_dep-2**	+/ville_dep-2
de	-	**+/ville_dep-1**	+/ville_dep-1	+/ville_dep-1
Grenoble	+/ville_dep	+/ville_dep	+/ville_dep	+/ville_dep
à	-	**+/ville_arr-1**	+/ville_arr-1	+/ville_arr-1
Clermont-Ferrand	+/ville_arr	+/ville_arr	+/ville_arr	+/ville_arr
en	-	**+/classe-1**	+/classe-1	+/classe-1
première	+/classe	+/classe	+/classe	+/classe
classe	-	-	-	**+/classe+1**

Figure 4: *Marqueurs de concepts automatiques. Les pré-marqueurs ont un ordre de p mots et les post-marqueurs de s. La 1ère colonne donne l'annotation manuelle en concepts de base. Les autres colonnes donnent le résultat de l'annotation automatique des marqueurs (en* **gras***).*

Table 3: *Taux d'erreurs de compréhension (%) et nombre total de concepts des modèles de compréhension, pour les différents ordres de marqueurs (p - s) sur les transcriptions automatiques de l'ensemble de test. La première colonne donne les résultats en n'utilisant aucun marqueur, la dernière donne les résultats avec les marqueurs manuels.*

Marqueurs	sans	1-0	2-0	3-0	1-1	**2-1**	3-1	manuels
Nb total concepts	116	222	322	413	316	**409**	493	148
Test	26,9	21,6	21,6	21,6	21,6	**19,8**	20,2	19,1

Le principe utilisé consiste à associer à chaque concept de base un ensemble de pré- et post-marqueurs. Ces marqueurs se distinguent par leur distance au concept. Un pré-marqueur (resp. post-marqueur) d'ordre p est attribué à chaque mot apparaissant p mots avant (resp. s suivant) un mot associé au concept de base correspondant dans l'ensemble d'apprentissage. Le mode est propagé, depuis le concept vers ses marqueurs. Un exemple d'annotations automatiques en marqueurs, obtenues pour différents ordres, est donné dans la figure 4.

La table 3 donne les résultats des expériences utilisant les transcriptions automatiques de l'ensemble de test, ainsi que le nombre total de concepts et de marqueurs de concepts manipulés par le modèle de compréhension. Lorsqu'aucun marqueur n'est utilisé, le taux d'erreurs est de 26,9%. Une diminution relative de 20% du taux d'erreurs est obtenue avec un pré-marqueur automatique d'ordre 1. L'accroissement seul de l'ordre des pré-marqueurs à $p = 2$ et $p = 3$ n'offre pas de réduction supplémentaire. Une amélioration des performances est obtenue avec l'introduction de post-marqueurs d'ordre 1. Le meilleur taux de compréhension est alors obtenu avec des pré- et des post-marqueurs d'ordres respectifs 2 et 1. L'apprentissage du modèle de compréhension à partir d'une annotation manuelle limitée aux seuls concepts de base et d'une annotation automatique en concepts marqueurs, permet de réduire le coût de l'annotation sans perte importante de performance : le taux d'erreurs de compréhension est de 19,8% contre 19,1% avec un modèle appris sur une annotation manuelle complète. Ce faible écart de performance est conservé en utilisant uniquement 2000 énoncés d'apprentissage (22,3% contre 21,5% avec les marqueurs manuels).

6. CONCLUSIONS

Les travaux présentés dans cet article constituent une première étape vers l'établissement d'un protocole précis et complet pour le développement et l'évaluation d'un module stochastique de compréhension pour les sytèmes de dialogue. Nos expériences sur une tâche de renseignements ferroviaires ont d'abord montré que les approches stochastique et par grammaire de cas offrent un niveau de performance comparable.

Les expériences montrent que le corpus d'apprentissage peut être réduit à quelques milliers d'énoncés sans perte importante de performance. Toutefois ce résultat indique aussi qu'il existe une marge d'amélioration des modèles impliquant une meilleure prise en compte des données supplémentaires.

Enfin, une procédure permettant l'annotation automatique des énoncés en terme de marqueurs de concepts a été présentée. Le coût de l'annotation manuelle est alors diminué par la réduction de 20% du nombre d'étiquettes prises en compte pour une diminution relative du taux de compréhension de seulement 3% .

7. REMERCIEMENTS

Nous remercions Sophie Rosset pour nous avoir fourni les résultats de l'approche par grammaire de cas.

BIBLIOGRAPHIE

[1] H. Bonneau-Maynard and F. Lefèvre. Investigating stochastic speech understanding. In *ASRU*, 2001.

[2] L. Lamel, S. Rosset, J.L. Gauvain, and S. Bennacef. The limsi arise system. *Speech Communication*, 31, 2000.

[3] W. Minker. *Compréhension Automatique de la Parole Spontanée*. PhD thesis, Université Paris XI, 1998.

[4] R. Pieraccini and E. Levin. A learning approach to natural language understanding. *NATO ASI Series Springer-Verlag*, 1993.

[5] G. Riccardi and A. Gorin. Stochastic language models for speech recognition and understanding. In *ICSLP*, Sidney, 1998.

[6] R. Schwartz and all. Hidden understanding models for statistical sentence understanding. In *ICASSP*, Munich, 1997.